Comparación de métodos de clasificación aplicados al mercado laboral costarricense

Isaura Gutiérrez Vargas[1], Aracelly Zárate Cordero[1], Wendy Zárate Cordero[1]
isaura.gutierrez@ucr.ac.cr, aracelly.zarate@ucr.ac.cr, wendy.zarate@ucr.ac.cr
Escuela de Estadística, Universidad de Costa Rica

RESUMEN

El desempleo es uno de los problemas económicos que más afecta a los costarricenses. Al primer trimestre de 2023, la tasa de desempleo se ubicó en 12% y se ha mantenido por encima de 11% en los últimos años. Predecir qué personas estarán empleadas es una tarea desafiante, pero es posible utilizar modelos de clasificación para identificar las variables más relevantes a nivel nacional que influyen en el acceso al mercado laboral. Con este objetivo, en este estudio se evaluaron 4 métodos de clasificación: árboles de decisión, bosques aleatorios, kNN-Vecinos más cercanos y regresión logística. Como resultado, se seleccionó el método de bagging con árboles de decisión debido a su menor tasa de error y a que obtiene el valor KS y AUC más alto. Se encontró que variables como el sexo, la edad, el estado conyugal y la educación del individuo son las más relevantes para predecir la condición laboral de los individuos.

PALABRAS CLAVE: métodos de clasificación, calibración, validación cruzada, aprendizaje supervisado, desempleo.

ABSTRACT

Unemployment is one of the economic problems that most affects Costa Ricans. In the first quarter of 2023, the unemployment rate stood at 12% and has remained above 11% in recent years. Predicting which individuals will be employed is a challenging task, but it is possible to use classification models to identify the most relevant variables at the national level that influence access to the labor market. With this objective, this study evaluated 4 classification methods: decision trees, random forests, kNN-Nearest Neighbors, and logistic regression. As a result, the bagging method with decision trees was selected due to its lower error rate and obtaining the highest KS and AUC values. It was found that variables such as gender, age, marital status, and individual education are the most relevant for predicting the employment status of individuals.

PALABRAS CLAVE: classification methods, calibration, crossed validation, unemployment.

[1] Estudiantes de Estadística, Universidad de Costa Rica

INTRODUCCIÓN

El desempleo es uno de los principales desafíos económicos que enfrenta Costa Rica. Las tasas de desempleo en el país han persistido por encima del 11%, con tasas de ocupación cercanas al 50% y un alto grado de informalidad laboral. De acuerdo con el Instituto Nacional de Estadística y Censos (INEC, 2023), los desempleados incluyen a aquellos sin empleo, pero que están disponibles para participar de la producción de bienes y servicios, y buscaron trabajo, pero no lo encontraron. Sin embargo, esta definición, aunque comúnmente utilizada por organismos de información estadística, solo abarca una parte del desempleo, excluyendo aspectos como la sobre educación o la subocupación (Castro Vincenzi et al., 2014).

Las razones por las que una persona decide no trabajar están influenciadas por muchos factores. Sin embargo, se pueden identificar características específicas que pueden influir en las oportunidades de acceso al mercado laboral. Según Robalino et al. (2021), en el caso del mercado laboral costarricense, los jóvenes y las mujeres son los grupos más afectados. Los investigadores señalan que los menores de 24 años presentan tasas de desempleo que duplican a las del resto de la población, mientras que las mujeres tienen tasas de desempleo más altas que los hombres durante todo el periodo comprendido entre 2010 y 2019.

De acuerdo con Blanco (2016), la menor participación de las mujeres en el mercado laboral se explica principalmente por una menor experiencia laboral y decisiones reproductivas. Esto es apoyado por la Organización Internacional del Trabajo (2014) que establece que las mujeres tienen mayores probabilidades de estar desempleadas que los hombres (Blanco, 2016). Esta situación no es particular para Costa Rica, sino que es un fenómeno ampliamente estudiado y que se observa en numerosos países. La Organización para la Cooperación y el Desarrollo Económicos (OCDE) destaca que, en países europeos con altas tasas de empleo, el desempleo femenino es sustancialmente más alto que el de los hombres.

La escolaridad de las personas también incide en la posibilidad de ser empleados. Según Robalino et al. (2021), las personas jóvenes y las personas sin educación primaria presentan altas tasas de inactividad en el país. Durante el periodo de 2010 a 2019, la tasa de desempleo de las personas de 15 a 24 años rondaba del 20% al 35%. La situación descrita se ha mantenido incluso después de la pandemia por Covid-19, porque los principales indicadores de empleo en el país se han devuelto a los niveles pre pandemia, pero sin presentar una mejora con respecto a la situación previa.

Considerando las características mencionadas anteriormente, podría parecer sencillo predecir cuáles personas en el país tienen mayores probabilidades de estar desempleadas. Históricamente, los modelos econométricos han sido utilizados para realizar estimaciones de desempleo. No obstante, en los últimos años, se ha recurrido a modelos de aprendizaje automático para obtener un entendimiento más profundo de las complejas relaciones que influyen en el desempleo.

En otras regiones, se han utilizado algoritmos de clasificación automática para evaluar la dinámica del mercado laboral (Kreiner, 2019). En un estudio realizado por Moumen et al. (2020), se emplearon técnicas como regresión logística, kNN y árboles de decisión para analizar la integración

de recién graduados universitarios en el mercado laboral. Además, en el campo de la economía, se ha explorado la combinación de estas técnicas con modelos econométricos tradicionales con el objetivo de mejorar la precisión de las predicciones de empleo (Chakraborty et al., 2021). Sin embargo, en el contexto costarricense, no se ha encontrado literatura que aborde el uso de estas técnicas para analizar datos de empleo nivel nacional.

Por tanto, el objetivo de esta investigación es encontrar la mejor regla de decisión, para determinar el estado de empleabilidad de una persona, y que además permita hacer clasificaciones futuras a nuevos individuos. Como objetivo específico, se busca caracterizar al mercado laboral costarricense en los años 2017 y 2022, para comprender la evolución de los principales indicadores y los cambios que ha experimentado producto de la pandemia por Covid-19.

METODOLOGÍA

Para llevar a cabo el análisis propuesto en este estudio, se utilizaron los datos de la Encuesta Continua de Empleo (ECE), los cuales fueron obtenidos a través del Programa Acelerado de Datos del INEC. Esta herramienta informática estandarizada proporciona acceso a microdatos y resultados estadísticos de las principales encuestas realizadas en el país (INEC, s.f). Dado que la ECE se lleva a cabo trimestralmente, se unieron los datos de las 4 encuestas realizadas en 2017 y también en 2022, para obtener bases de datos anuales. De esta manera, se dispone de un total de 44.879 observaciones para el año 2017 y 40.138 observaciones para el año 2022.

La ECE se compone de 388 variables, pero para este estudio únicamente se seleccionan 12 variables independientes, las cuales se describen en el Cuadro 1. Por otro lado, la variable respuesta en este estudio es binaria y representa si una persona está desempleada o no. Para algunas variables independientes, como el Estado Conyugal, se decidió combinar varias categorías de respuesta, con el propósito de crear categorías que contengan a un mayor número de individuos.

Cuadro 1
Variables independientes utilizadas en el análisis

Variable	Descripción	Niveles
Sexo	Característica biológica que distingue a las personas entrevistadas.	1. Hombre 2. Mujer
Edad	Hace referencia a los años de vida cumplidos al momento de la entrevista, independientemente del tiempo que deba transcurrir para que vuelva a cumplir años.	Variable continua de 0 a 99.
Estado Conyugal	El estado conyugal se comprende como la “situación actual” en que se encuentra la persona sin considerar la legalidad que haya.	1. Casado 2. Soltero 3. Unión Libre
Educación Asiste	Nivel educativo en el que se encuentra	1. Asiste

	asistiendo la persona en el momento de la entrevista.	2. No asiste
Educación Nivel Grado	Último grado o año aprobado en relación al nivel educativo de la persona, en la educación formal.	1. Ninguno 2. Secundaria o menos 3. Estudio superior
Educación No Regular Asiste	Aparte de la educación regular, ha recibido algún curso u otro tipo de formación del que tenga título o certificación.	1. Asiste 2. No asiste
Idioma	Habla, lee y escribe fluidamente algún otro idioma aparte de su lengua materna.	1. No 2. Sí
Región	Región socioeconómica de residencia.	1. Brunca 2. Central 3. Chorotega 4. Huetar Caribe 5. Huetar Norte 6. Pacífico Norte
Zona	Tipo de zona de residencia.	1. Rural 2. Urbana
País Nacimiento	País de nacimiento de las personas, según dónde vivía la mamá cuando nació.	1. Costa Rica 2. Otro
Permanencia País	Es residente de Costa Rica.	1. No residente 2. Residente

Para clasificar a los empleados y desempleados del país se ponen a prueba 4 métodos de clasificación: árboles de decisión, bosques aleatorios, k-vecinos más cercanos y regresión logística. Adicionalmente, se utilizó el método de ensamblaje bagging para las técnicas de regresión logística y árboles de decisión, lo cual consiste en entrenar el modelo con diferentes sets de datos lo cual ayuda a mejorar la estabilidad y a reducir la varianza del modelo (Reddy et al., 2018)

Cada uno de estos métodos se somete a una etapa de calibración inicial, para determinar los valores más adecuados de los parámetros. Para realizar este paso se aplica validación cruzada para cada valor posible de los parámetros a probar, obteniendo las medias de los indicadores de desempeño. Por último, los 4 métodos se someten a validación cruzada para determinar el más óptimo para clasificar a las personas según su condición de actividad.

La primera técnica por utilizar son los árboles de decisión, los cuales sirven para tomar decisiones basadas en un conjunto de variables predictoras. De acuerdo con Liu & Liu (2022), los árboles de decisión presentan ciertas ventajas como que los resultados son fáciles de interpretar y

entender. Los resultados finales de la clasificación se describen en una forma de IFELSE que está en línea con el pensamiento lógico de las personas al analizar problemas. Además, los cálculos necesarios son pocos y por tanto la eficiencia del método es alta. Para esta técnica se debe calibrar la profundidad y la complejidad del árbol; la profundidad consiste en el número de niveles que tiene el árbol mientras que la complejidad se relaciona con la cantidad de nodos y ramas que tiene (Buhrman & de Wolf, 2002).

El clasificador de bosques aleatorios es una de las técnicas de aprendizaje en conjunto más exitosas que ha demostrado ser muy popular y poderosa en el reconocimiento de patrones y el aprendizaje automático para clasificación de alta dimensionalidad y problemas sesgados (Azar et al., 2014). Este consiste en una combinación de varios árboles de decisión. Los bosques aleatorios se pueden construir muestreando aleatoriamente un subconjunto de características. Los parámetros a calibrar en este método es el número de árboles con los que se tomará la decisión final y el número de variables aleatorias.

La tercera técnica es el uso de k vecinos más cercanos, también conocida como kNN. Este es un algoritmo simple que se utiliza para predecir las etiquetas de los puntos de datos de prueba después de entrenar con una muestra de datos. Esta técnica encuentra un grupo de k objetos en el conjunto de entrenamiento que están más cerca del objeto de prueba y basa la asignación de una etiqueta en la predominancia de una clase particular en este vecindario (Wu et al., 2008). Para esta técnica, se debe calibrar el número de vecinos (k) óptimo. Normalmente kNN utiliza la distancia euclídea, pero para el tratamiento de la ECE se usa la distancia de Gower por la presencia de variables numéricas y categóricas.

Por último, se implementará el modelo logístico binomial. En casos de clasificación, la regresión logística devuelve la probabilidad de que ocurra un evento particular, estimada utilizando un modelo logístico (de Menezes et al., 2017). Para el modelo logístico, se calibra el umbral de probabilidad utilizado como punto de corte para las predicciones del modelo. Además, tanto para el modelo logístico como para los árboles de decisión se realiza un proceso de selección de variables previo a la calibración del resto de parámetros.

Es importante mencionar que la base de datos que se trabaja presenta un problema de desbalance en la variable respuesta, puesto que hay más empleados que desempleados. Esto representa un reto dado que los modelos de clasificación operan bajo el supuesto de que los datos de entrenamiento tienen clases balanceadas en la variable respuesta, por lo que una variable desbalanceada no daría resultados satisfactorios (Sun & Chen, 2021).

De acuerdo con Ali et al. (2019), la mayoría de las veces las clases mayoritarias sesgan los clasificadores a favor de sí mismas y el clasificador presenta de manera deficiente las tasas de clasificación de las clases minoritarias; eventualmente, un clasificador se dirige completamente como clase mayoritaria e ignora la clase minoritaria. Para resolver este problema, se propone el uso de ponderadores para darle una importancia relativa mayor a la clase con menos observaciones (Prati et al., 2009).

En la etapa de validación, se consideraron varios indicadores de desempeño para evaluar el rendimiento del modelo. Estos indicadores incluyeron los falsos negativos (FN), falsos positivos (FP),

el área bajo la curva ROC (AUC) y el estadístico Kolmogorov-Smirnov (KS). Un modelo satisfactorio tendría un AUC con valores entre 0.6 y 0.9, un KS entre 0.2 y 0.7, y un error, FN y FP bajos. Con base en estos parámetros, se escogerá la técnica de clasificación que dé los resultados más satisfactorios.

Para realizar el análisis se utiliza el software estadístico R, versión 4.3.1 (R Core Team, 2022), para realizar el análisis. Además, se emplearon las siguientes librerías: rattle (Williams G, 2011), caret (Kuhn, 2022), rocr (Sing T, Sander O, Beerenwinkel N y Lengauer T, 2005) para realizar las diferentes técnicas. En particular, para la técnica de árboles de decisión se utilizó la librería rpart (Therneau y Atkinson, 2022), para los bosques aleatorios se empleó randomForest (Liaw y Wiener, 2002). Por último, se utilizó el paquete ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York (Wickham, H. 2016) para la creación de gráficos para mostrar los criterios de validación y la estructura de los datos.

RESULTADOS

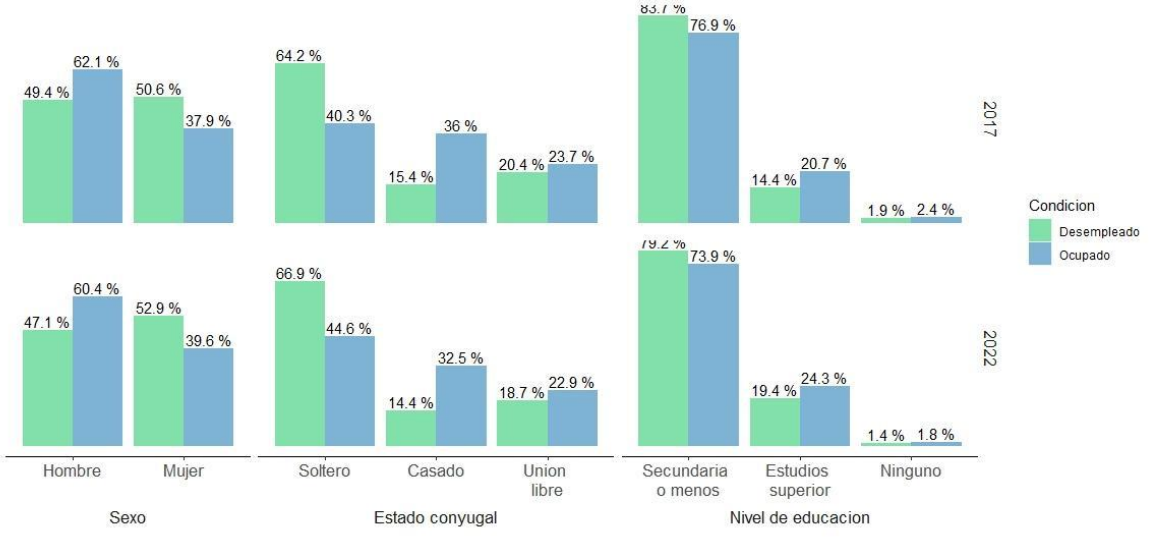
Para comprender el comportamiento inicial de los datos, es importante realizar una inspección inicial de las variables de interés y su relación con la variable dependiente. En este contexto, se presenta la Figura 1, la distribución porcentual de algunas de las variables relevantes, como el sexo, estado conyugal y nivel de educación del individuo, para los años 2017 y 2022.

Es importante destacar que, en el año 2017, solo el 10.72% (4811 personas) de las personas entrevistadas afirmaron estar desempleadas, mientras que el 89.28% (40068 personas) aseguraron estar ocupadas. De manera similar, en el 2022, el 12.18% (4889 personas) mencionaron estar desempleadas, y el restante 87.82% (35248 personas) se encontraban ocupadas.

Al analizar los datos para ambos años, se observa un comportamiento similar en las variables estudiadas. Por ejemplo, se destaca que las personas casadas presentan los niveles más bajos de desempleo, con un 15.4% en 2017 y un 14.4% en 2022. Por otra parte, se observa que las mujeres muestran los mayores porcentajes de desempleo en ambos años. Un aspecto relevante a destacar es que la fuerza laboral está compuesta principalmente por personas con educación secundaria o inferior. Para el año 2017, este grupo representó aproximadamente el 79.9% de la fuerza laboral, mientras que en el año 2022 fue del 73.9%. La distribución porcentual de las restantes ocho variables se muestra en la Cuadro 1.a (en anexos).

Figura 1

Distribución porcentual de las variables sexo, estado conyugal, nivel de educación, por año, según la condición laboral.



Como se menciona en la metodología, se ejecutaron diversos métodos de clasificación con el objetivo de obtener el método más efectivo para clasificar a las personas según la condición laboral. Para lograrlo, se inició con un proceso de selección de variables adecuadas tanto para el modelo logístico como para el árbol de decisión. Este proceso consistió en eliminar una variable a la vez del modelo correspondiente y posteriormente evaluar su rendimiento. Los resultados de esta evaluación se encuentran detallados en la Cuadro 2 para las variables utilizadas para el año 2017 y en la Cuadro 3 para las variables utilizadas para el año 2022.

Cuadro 2

Indicadores de desempeño después de eliminar una variable a la vez para los datos del 2017

Sin variable	Error	FP	FN	AUC
Sexo	34.03	32.63	34.21	72.40
Edad	35.79	36.06	35.76	70.22
Estado conyugal	34.74	29.39	35.40	72.50
Asiste a educación	34.31	29.70	34.89	73.67
Nivel educación	34.68	29.60	35.31	73.24
Asiste a educación no regular	33.81	30.40	34.23	73.59
Idioma	33.68	30.40	34.08	73.71
Región	34.49	30.30	35.01	72.53
Zona	33.88	30.40	34.31	73.75
País nacimiento	33.91	30.10	34.38	73.76
Permanencia país	33.95	30.40	34.40	73.76

Cuadro 3

Indicadores de desempeño después de eliminar una variable a la vez para los datos del 2022

Sin variable	Error	FP	FN	AUC
Sexo	34.82	34.79	34.82	69.58
Edad	40.65	32.87	41.68	67.06
Estado conyugal	36.07	31.06	36.66	70.35
Asiste a educación	35.94	31.80	36.49	70.66
Nivel educación	35.85	32.12	36.35	70.24
Asiste a educación no regular	35.01	32.66	35.32	70.95
Idioma	34.74	32.44	35.05	70.94

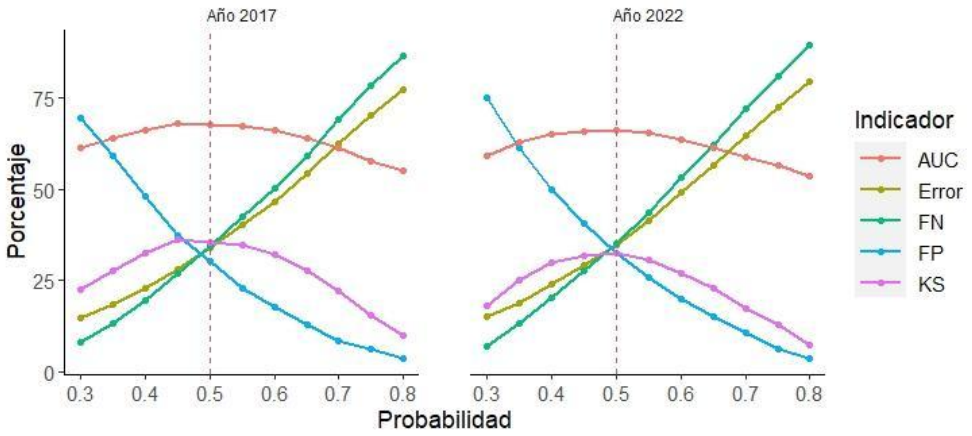
Región	34.63	32.55	34.10	70.86
Zona	34.74	31.91	35.12	70.94
País nacimiento	34.68	32.34	34.99	71.01
Permanencia país	34.71	32.34	35.02	71.00

Se observó que, al excluir una variable cada vez del modelo de clasificación, no se encontraron cambios importantes en los indicadores de desempeño. Esto indica que las 11 variables utilizadas son útiles para entrenar los diferentes modelos, por lo que se optó por incluir todas las variables en los análisis subsiguientes.

Posteriormente, se llevó a cabo la calibración del umbral de probabilidad utilizado como punto de corte para las predicciones del modelo logístico. La Figura 2 muestra que cuando se utilizan probabilidades inferiores a 0.5, se obtiene un alto porcentaje de falsos positivos (FP) y un bajo porcentaje de falsos negativos (FN). En contraste, al utilizar probabilidades mayores a 0.5, se reduce la cantidad de falsos positivos, pero aumenta de manera importante la cantidad de falsos negativos. Por lo que, de acuerdo con estos resultados, se decidió utilizar un umbral de probabilidad de 0.5, representado por la línea punteada roja. Con este umbral se obtiene el mayor valor para el indicador AUC y en el indicador KS, tanto para el año 2017 como para el año 2022.

Figura 2

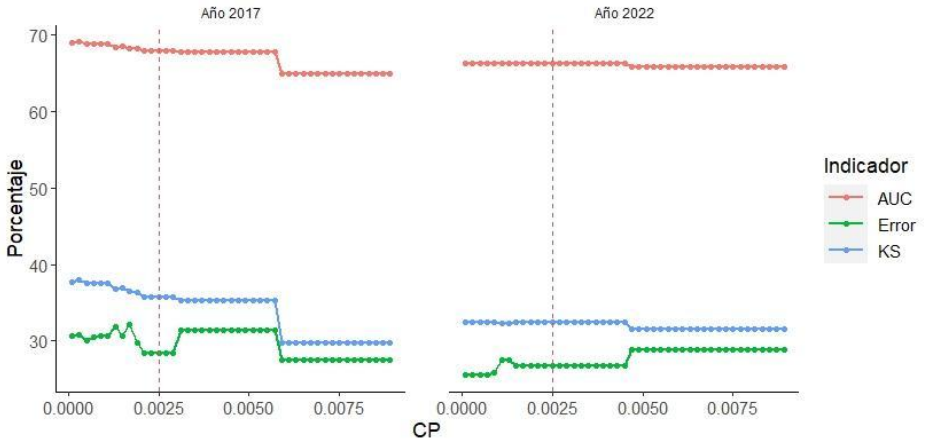
Calibración del parámetro de probabilidad para el modelo de clasificación logístico.



En el modelo de árboles de decisión, se llevó a cabo la calibración del parámetro de complejidad (CP) utilizando los indicadores de AUC, KS y error de clasificación. Se probaron varios valores de CP y se evaluó su impacto en los indicadores. Los resultados se presentan en la Figura 3, para ambos años, se seleccionó el valor de complejidad (0,0025) ya que generó no solo el AUC y el indicador KS más altos, sino que también resultó en el menor error de clasificación

Figura 3

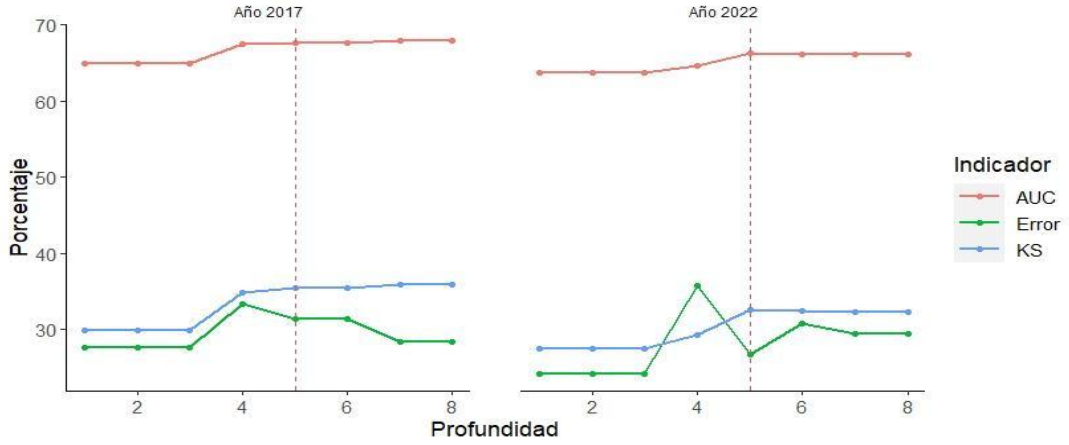
Calibración del parámetro de complejidad (CP) para el método de árboles de decisión.



Posteriormente, se realizó la calibración del parámetro de profundidad del árbol utilizando los indicadores de error de clasificación, AUC y KS. Se probaron valores en un rango de 1 a 8 para ambos años, manteniendo constante el parámetro de complejidad previamente seleccionado. Los resultados de esta calibración se presentan en la Figura 4.

Figura 4

Calibración del parámetro de profundidad para el método de árboles de decisión.

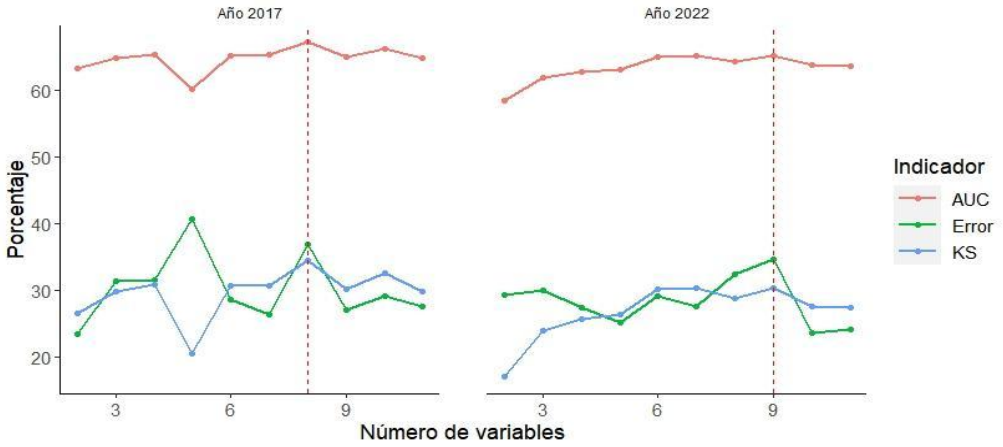


Con base en los resultados obtenidos, se seleccionó un valor de 5 para el parámetro de profundidad tanto en el año 2017 como en el año 2022. Esta elección se basó en el hecho de que con esta profundidad se logran valores altos en los indicadores AUC y KS, mientras que se disminuye el error de clasificación. Finalmente, se procedió a la calibración del modelo de Bosques Aleatorios, en particular, se ajustó el número de variables aleatorias utilizando los indicadores de AUC, KS y error de clasificación. Los resultados de esta calibración se presentan en la Figura 5.

Utilizando el criterio de AUC y KS, se determinó que el número adecuado de variables aleatorias es de ocho para el año 2017. Mientras que, para los datos del año 2022, se estableció que el número óptimo de variables es de nueve.

Figura 5

Calibración del parámetro de número de variables aleatorias para el método de bosques aleatorios



A su vez, para el modelo de bosques aleatorios se calibró el número de árboles que genera el modelo para realizar la predicción final. Debido a limitaciones computacionales se probaron dos valores para este parámetro: 3 y 5 árboles. Los resultados se muestran en el Cuadro 4.

Se observó que, a medida que se utilizan más árboles (5) para la predicción final, el indicador AUC aumenta en ambos años. En el caso del año 2017, se observó un aumento en el error de clasificación, pero también una disminución en los falsos positivos. Este aumento en el error podría atribuirse al hecho de que el modelo mejora la predicción de los desempleados, pero aumenta la cantidad de falsos negativos (individuos ocupados clasificados incorrectamente como desempleados).

Para el año 2022, se encontró una disminución considerable en el error de clasificación al utilizar 5 árboles para la predicción final. Aunque hubo un aumento en el porcentaje de falsos positivos, se observó una disminución considerable en los falsos negativos. Por tanto, se decidió fijar el parámetro de cantidad de árboles en 5 para ambos años.

Cuadro 4

Indicadores de desempeño según el número de árboles utilizados en la técnica de Bosques aleatorios por año

Indicador	2017		2022	
	3	5	3	5
Error	32.70	34.00	47.29	35.92
FP	36.87	29.80	24.97	34.47

FN	32.18	34.51	50.24	36.11
AUC	65.47	67.84	62.39	64.71
KS	30.95	35.69	24.79	29.42

A continuación, en la Figura 6 y Figura 7 presentadas en anexos, se muestra la representación gráfica de los árboles de decisión construidos para los años 2017 y 2022, respectivamente, utilizando las 11 variables mencionadas previamente y los parámetros seleccionados. Al comparar los árboles, se pueden identificar cambios en las variables más relevantes o en la estructura general del modelo. Esto puede proporcionar información valiosa sobre cómo las variables han influido en las decisiones del modelo a lo largo del tiempo.

En este caso particular, se observa una diferencia en las variables utilizadas en los dos modelos. Ambos árboles de decisión comienzan con una regla basada en la edad, clasificando a las personas mayores de 30 años en el árbol correspondiente al año 2017 y a las personas mayores de 29 años en el árbol correspondiente al año 2022. Sin embargo, a partir de ese punto, las variables utilizadas para la clasificación difieren.

En el modelo del año 2022, después de la clasificación por edad, se incorporan variables como el estado conyugal, la región de residencia, nuevamente la edad y, por último, el nivel de educación. En contraste, en el modelo del año 2017, las personas mayores de 29 años se clasifican según su estado conyugal, nivel de educación, la edad y, finalmente, el sexo.

Una vez calibrados los métodos se procedió a realizar la validación cruzada para escoger cuál es el método ideal para clasificar a los individuos. Asimismo, se incluyeron las técnicas de bagging logístico, bagging con árboles de decisión, bosques aleatorios y k-vecinos más cercanos, siguiendo los parámetros establecidos anteriormente.

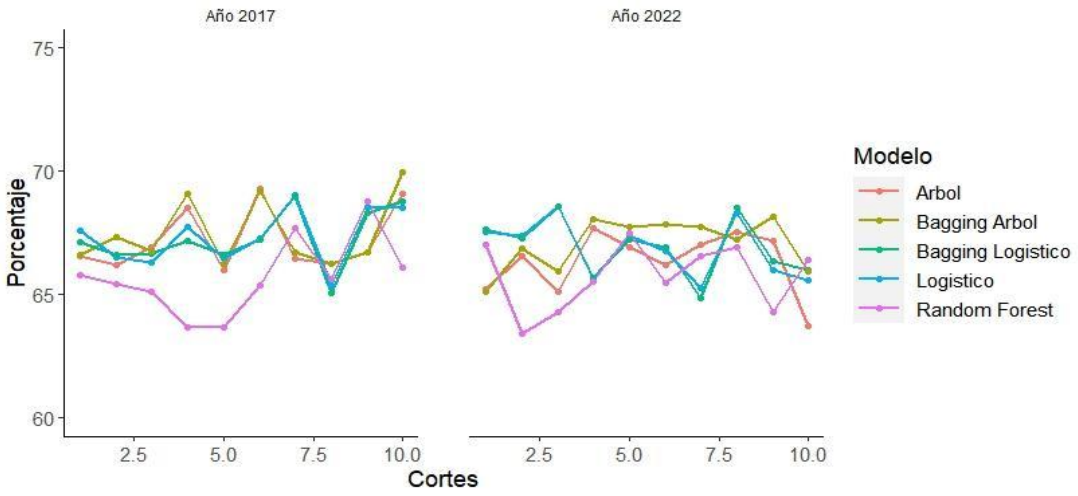
En la Figura 8 se presentan los valores obtenidos en cada pliegue para los diferentes modelos, utilizando el criterio AUC como medida de rendimiento. Es importante mencionar que la técnica de k vecinos más cercanos no es comparable con el resto de modelos, ya que debido a la naturaleza de la técnica se recurrió a utilizar una base de datos más pequeña, que solo incluye información sobre la ECE para el primer trimestre del 2017 y el primer trimestre de 2022.

Al analizar el rendimiento de los modelos en los 10 pliegues realizados, se observa que, en el caso del año 2017, el modelo de árbol de decisión es muy similar al obtenido aplicando la técnica de bagging con este mismo modelo, de forma similar, el modelo logístico da el mismo resultado que aplicar bagging logístico, estos cuatro modelos tienen un rendimiento similar, sin embargo el modelo de bosques aleatorios muestra un rendimiento relativamente más débil en comparación con los otros modelos evaluados.

De igual forma, para los datos del año 2022, el modelo logístico y bagging logístico brindan resultados muy parecidos, en este caso se observa que el modelo de bagging con árboles de decisión es un poco mejor que el modelo de árboles de decisión, además, el modelo de bosques aleatorios tiene un rendimiento similar en comparación con el resto de modelos.

Figura 8

Indicador de desempeño AUC, por año, según modelo de clasificación

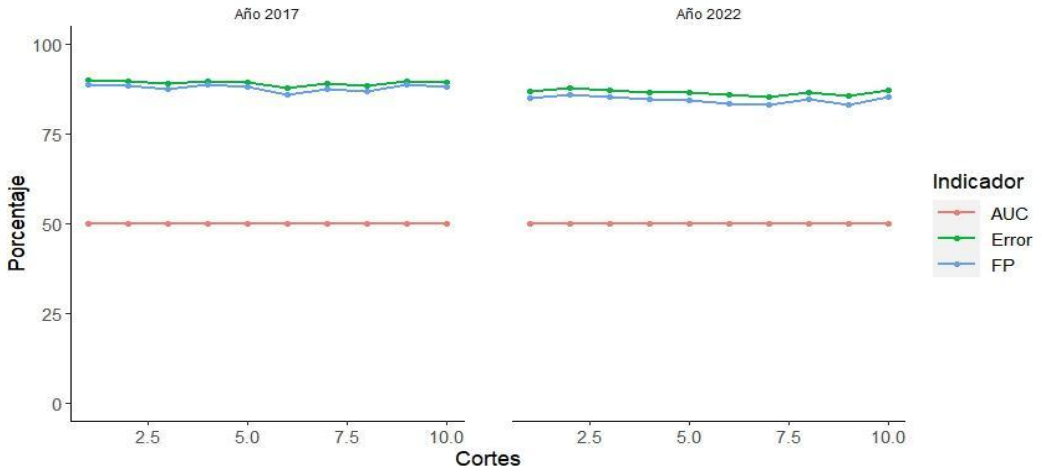


En relación con los resultados obtenidos al emplear el método de k vecinos más cercano para realizar la clasificación de los individuos desempleados y ocupados, se presentan en la Figura 9. En la misma es evidente que tanto el error de clasificación como los falsos positivos tienen valores muy altos, aproximadamente alrededor del 90%. Esto significa que el modelo está cometiendo errores al clasificar incorrectamente a los individuos desempleados como ocupados. Por otro lado, no se están cometiendo errores en términos de falsos negativos, lo que indica que los individuos ocupados no están siendo clasificados erróneamente como desempleados.

El pésimo rendimiento del modelo al clasificar a los desempleados es notable, y esto puede explicarse debido a que la categoría de desempleados es minoritaria en comparación con la cantidad de individuos ocupados.

Figura 9

Indicadores de desempeño por año, para el método de K vecinos más cercanos



Finalmente, en el Cuadro 5 se presentan los promedios de los indicadores considerando los valores obtenidos en cada uno de los pliegues. Los promedios proporcionan una visión general del desempeño promedio de los modelos en términos de los indicadores evaluados. Al calcular los promedios de los indicadores, se tiene en cuenta la variabilidad en los resultados obtenidos en cada pliegue y se obtiene una medida más estable del rendimiento global.

Cuadro 5

Indicadores de desempeño por año, según la técnica de clasificación.

	2017					2022				
Modelo	**Error**	**FP**	**FN**	**AUC**	**KS**	**Error**	**FP**	**FN**	**AUC**	**KS**
Logístico	34.01	30.93	34.39	67.34	34.68	34.46	31.41	34.89	66.85	33.71
Árbol	30.69	35.47	30.12	67.20	34.40	28.53	40.46	26.89	66.32	32.65
Bagging Logístico	34.01	31.12	34.36	67.26	34.52	34.32	31.46	34.72	66.91	33.82
Bagging Árbol	29.51	36.13	28.71	**67.58**	35.16	28.15	39.24	26.62	**67.07**	34.14
Random Forest	28.60	41.46	27.06	65.74	31.47	30.30	39.43	29.06	65.75	31.51

Utilizando el criterio de AUC, se determinó que para ambos años el modelo con el mayor AUC es la técnica de bagging aplicada a un árbol de decisión. En el año 2017, este modelo presenta uno de los puntajes más bajos en términos de error de clasificación, lo que implica que en promedio el 29,51% de los individuos fueron clasificados incorrectamente. De manera similar, en el año 2022 utilizando el modelo de bagging con árboles de decisión, el 28,15% de los individuos fueron mal clasificados.

En cuanto a los falsos positivos en el año 2017, representan el 36,13%, lo que indica que este porcentaje de individuos fue clasificado como ocupados, pero en realidad estaban desempleados. Por otro lado, se encontró un 28,71% de falsos negativos en ese mismo año, lo que significa que este porcentaje de individuos fue clasificado como desempleados, pero en realidad estaban ocupados.
En el caso del AUC, representa el porcentaje de clasificación adecuada de los individuos que están ocupados y es del 67,58%, lo cual es más alto que el porcentaje de clasificación errónea de los individuos desempleados.

Para el año 2022, el 39,24% de los individuos clasificados como ocupados en realidad estaban desempleados, mientras que el 26,62% de los individuos clasificados como desempleados estaban ocupados. El AUC de este modelo en este año indica que el porcentaje de clasificación adecuada de los individuos ocupados es del 67,07%, también más alto que el porcentaje de clasificación errónea de los individuos desempleados.

CONCLUSIONES

El desempleo en Costa Rica es uno de los principales desafíos económicos del país, con tasas persistentemente altas por encima del 11% y una alta informalidad laboral. Además, el desempleo afecta de manera desproporcionada a ciertos grupos, como a las personas más jóvenes, las mujeres y a las personas inmigrantes. Dentro del análisis de datos que se realizó, destaca el hecho de que los principales indicadores del mercado laboral no presentan variaciones entre los años 2017 y 2022, lo que demuestra que el mercado laboral costarricense regresó a los niveles pre pandemia sin presentarse una mejora en los indicadores para los grupos más excluidos del mercado.

En cuanto al mejor método de clasificación, después de comparar los indicadores de desempeño, se determinó que el bagging con árboles de decisión da el valor de AUC y KS más alto y disminuye el error. Por el contrario, el método con peores resultados es el uso de árboles de decisión y regresión logística, puesto que presentan los mayores porcentajes de error para el año 2017 y 2022, respectivamente. Sin embargo, la técnica de árboles de decisión presenta indicadores de desempeño similares a los obtenidos por bagging y su costo computacional es inferior. Por lo que se recomendaría esta técnica como la más efectiva en relación con su complejidad.

Los métodos utilizados presentan un porcentaje de Falsos Positivos que ronda de 30.93%-41.46% y de Falsos Negativos de 26.62%-34.89%, a pesar de que estos porcentajes son más altos de lo que se desea, se debe tomar en cuenta que la Encuesta Continua de Empleo es representativa de todo el país y por tanto existe una alta heterogeneidad en los individuos que componen la muestra. Tomando en consideración únicamente 11 variables independientes, se logra un indicador AUC y KS satisfactorio, lo cual es un resultado favorable hacia la utilización de algoritmos de clasificación a nivel nacional.

Entre las limitaciones que se enfrentaron en este estudio, se destaca la falta de variables relacionadas con la condición económica del individuo. Robalino et al. (2021) destaca que los individuos del primer quintil de ingresos presentan tasas de desempleo inferiores al resto de la población, por lo que incluir alguna variable relacionada al ingreso podría ayudar a disminuir el porcentaje de error de las diferentes técnicas.

Otra limitación se relaciona con la capacidad computacional requerida para llevar a cabo los diferentes métodos de clasificación. En el caso del kNN, no fue posible aplicar esta técnica al conjunto completo de datos debido a la duración prolongada de ejecución de la técnica. Además, en el método de bagging y bosques aleatorios, no se pudo aumentar la cantidad de árboles de decisión más allá de 5 debido a restricciones computacionales.

REFERENCIAS

Azar, A. T., Elshazly, H. I., Hassanien, A. E., & Elkorany, A. M. (2014). A random forest classifier for lymph diseases. *Computer Methods and Programs in Biomedicine*, *113*(2), 465-473. https://doi.org/10.1016/j.cmpb.2013.11.004

Azmat, G., Güell, M., & Manning, A. (2006). Gender Gaps in Unemployment Rates in OECD Countries. *Journal of Labor Economics*, *24*(1), 1-37. https://doi.org/10.1086/497817

Ali, H., Salleh, M. N. M., Saedudin, R., Hussain, K., & Mushtaq, M. F. (2019). Imbalance class problems in data mining: A review. *Indonesian Journal of Electrical Engineering and Computer Science*, 14(3), Article 3. https://doi.org/10.11591/ijeecs.v14.i3.pp1552-1563

Blanco, L. C. (2016). Relación entre la segregación de género en las disciplinas de estudio universitarias y el empleo de las personas recién graduadas en Costa Rica. *Working Papers*, Article 201604. https://ideas.repec.org//p/fcr/wpaper/201604.html

Buhrman, H., & de Wolf, R. (2002). Complexity measures and decision tree complexity: A survey. *Theoretical Computer Science*, *288*(1), 21-43. https://doi.org/10.1016/S0304-3975(01)00144-X

Castro Vincenzi, J. M., Garita Garita, J., & Odio Zúñiga, M. (2014). Análisis sobre la dinámica de transición y duración del desempleo en Costa Rica. *Revista De Ciencias Económicas*, *32*(2), 39–64. https://doi.org/10.15517/rce.v32i2.17251

Celbiş, M. G. (2022). Unemployment in Rural Europe: A Machine Learning Perspective. Applied Spatial Analysis and Policy. https://doi.org/10.1007/s12061-022-09464-0

Chakraborty, T., Chakraborty, A. K., Biswas, M., Banerjee, S., & Bhattacharya, S. (2021). Unemployment Rate Forecasting: A Hybrid Approach. *Computational Economics*, *57*(1), 183-201. https://doi.org/10.1007/s10614-020-10040-2

De Menezes, F. S., Liska, G. R., Cirillo, M. A., & Vivanco, M. J. F. (2017). Data classification with binary response through the Boosting algorithm and logistic regression. *Expert Systems with Applications*, *69*, 62-73. https://doi.org/10.1016/j.eswa.2016.08.014

Gender Gaps in Unemployment Rates in OECD Countries | Journal of Labor Economics: Vol 24, No 1. (s. f.). https://www.journals.uchicago.edu/doi/full/10.1086/497817?casa_token=iRwaqaC1A-4AAAAA%3Aq2xg6aJwEZGAuIWnhFzC13bll8c_6pk0Jes6gOsm6lQ98vuEkNzAIg58_we0tDizXyCZGMI9tsw5k

INEC. (s. f.). *Catálogo Central de Datos*. Recuperado 10 de julio de 2023, de http://sistemas.inec.cr/pad5/index.php/catalog/central

Kreiner, A. (2019). Can Machine Learning on Economic Data Better Forecast the Unemployment Rate? *Honors Papers*. https://digitalcommons.oberlin.edu/honors/126

Kuhn, M. (2022). *_caret: Classification and Regression Training_. R package. V*ersión 6.0-92, <https://CRAN.R-project.org/package=caret>.

Robalino, J., Laura Cristina Blanco, L. C., Paredes, S., Mayorga, B., & Córdoba, D. (2021). *Informe sobre la evolución del mercado laboral en Costa Rica. Tendencias 2010-2019* (Working Papers N.º 202102). Universidad de Costa Rica. https://EconPapers.repec.org/RePEc:fcr:wpaper:202102

Liaw, A., y Wiener, M. (2002). *Classification and Regression by randomForest. R News.* 2(3), 18--22.

Liu, H., & Liu, J. (2022). Female Employment Data Analysis Based on Decision Tree Algorithm and Association Rule Analysis Method. *Scientific Programming, 2022,* e8994349. https://doi.org/10.1155/2022/8994349

Moumen, A., Bouchama, E.H., & Idrissi, Y.E. (2020). Data mining techniques for employability: Systematic literature review*. 2020 IEEE 2nd International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS)*, 1-5.

Prati, R.C., Batista, G.E., & Monard, M.C. (2009). Data mining with imbalanced class distributions: concepts and method*s. Indian International Conference on Artificial Intelligence.*

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Reddy, U. S., Thota, A., & Dharun, A. (2018). *Machine Learning Techniques for Stress Prediction in* Working Employees. 1-4. https://doi.org/10.1109/ICCIC.2018.8782395

Sing, T., Sander, O., Beerenwinkel, N., Lengauer, T. (2005). *ROCR: visualizing classifier performance in R_Bioinformatics_*, 21(20), 7881. http://rocr.bioinf.mpi-sb.mpg.de.

Sun, B., & Chen, H. (2021). A Survey of k Nearest Neighbor Algorithms for Solving the Class Imbalanced Problem. *Wireless Communications and Mobile Computing*, *2021*, 1-12. https://doi.org/10.1155/2021/5520990

Therneau, T., Atkinson, B., (2022). *rpart: Recursive Partitioning and Regression Trees_. R package version 4.1.16,* https://CRAN.R-project.org/package=rpart.

Wickham, H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.

Williams, G. (2011), *Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery, Use R!*, Springer.

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z.-H., Steinbach, M., Hand, D. J., & Steinberg, D. (2008). Top 10 algorithms in data

mining. *Knowledge and Information Systems*, 14(1), 1-37. https://doi.org/10.1007/s10115-007-0114-2

ANEXOS

Cuadro 1.a

Distribución porcentual de las variables categóricas por año, según la condición laboral del individuo

	2017		2022	
	Ocupados	**Desempleados**	**Ocupados**	**Desempleados**
Asiste a educación	9.2	19.0	9.0	20.5
No asiste a educación	90.8	81.0	91.0	79.5
Asiste a educación no regular	34.4	30.5	7.1	6.6
No asiste a educación no regular	65.6	69.5	92.9	93.4
No posee segundo idioma	93.7	95.7	92.5	94.1
Posee segundo idioma	6.3	4.3	7.5	5.9
Zona Rural	43.4	41.9	44.2	42.0
Zona Urbana	56.6	58.1	55.8	58.0
Nacional	90.3	90.7	89.9	90.2
Extranjero	9.7	9.3	10.1	9.8
No residente	9.6	8.9	10.0	9.5
Residente	90.4	91.1	90.0	90.5

Figura 6

Árbol de decisión para el año 2017

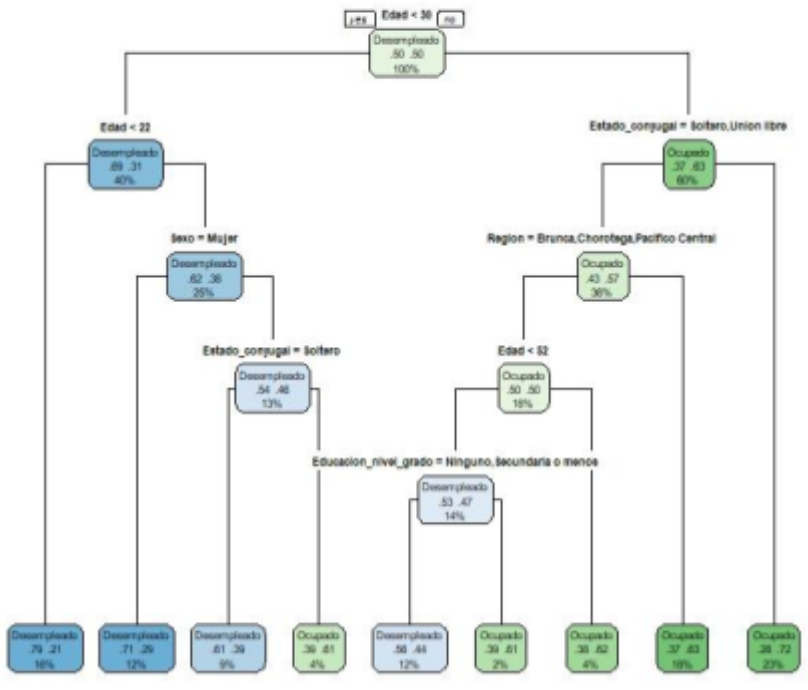


Figura 7

Árbol de decisión para el año 2022